The influence of similarity between concepts in evolving biomedical ontologies for mapping adaptation

Julio Cesar DOS REIS[a,b,1], Duy DINH[a], Cédric PRUSKI[a], Marcos DA SILVEIRA[a], Chantal REYNAUD-DELAÎTRE[b]
[a]*Public Research Centre Henri Tudor, Luxembourg*
[b]*LRI, University of Paris-Sud XI, France*

Abstract. Biomedical ontologies continuously evolve and impact associated mappings. This demands to adapt ontology mappings to maintain them up-to-date. This article studies whether similarity calculated between values of concept attributes issued from successive ontology versions plays a role in deciding mapping adaptation actions. We empirically analyse the evolution of official mappings established between large biomedical ontologies. The results point out the relevance of this factor for mapping adaptation.

Keywords. mapping adaptation, mapping evolution, mapping maintenance, similarity metrics, ontology alignment, ontology evolution, ontology versions

1. Introduction

The needs of exchanging and retrieving data between biomedical systems have increasingly become relevant. Ontology mappings play a key role in enabling semantic interoperability in this context [1]. They interconnect concepts of domain-related ontologies allowing systems to interpret data annotated with different ontologies. The huge size and existing intersections between ontologies force linking them through mappings. However, to remain useful and reflect the most up-to-date knowledge of the domain, these ontologies evolve and new versions are periodically released. This potentially impacts existing mappings demanding methods to ensure, as automatic as possible, their semantic consistency over time.

We have designed the *DyKOSMap* framework to adapt ontology mappings through a set of ***Mapping Adaptation Actions (MAAs)*** [2]. Our approach aims at deciding which action to apply when some ontology change affects the source concept of a correspondence. Existing tools enable to calculate simple and complex ontology changes given two successive ontology versions [3]. Indeed, various factors may influence the action decision for each correspondence individually [2]. We have studied how different types of ontology changes correlate with the adaptation actions [4][5][6]. Nevertheless, the complete understanding of this phenomenon demands further studies inquiring other factors to take into account in the mapping adaptation process.

This article reports on the influence of the similarity relatedness for mapping adaptation. We calculate the similarity between textual statements from source

[1] Corresponding Author.

concepts of mappings with textual statements of parents, children and sibling concepts issued from a new ontology version. These statements are attribute values characterizing the concepts. For instance, an attribute a_i, of type name, contains the value *"cardio vascular disease"*. We investigate whether the MAAs correlate with the behaviour of similarity values observed. We hypothesize that the similarity aspect stands for an element which may help deciding the adequate actions to adapt mappings.

2. Methods

An ontology *O* consists of a set of concepts interrelated by directed relationships. Each concept c_i^j at time *j* has a unique identifier, and a set of attributes that characterizes concepts where $ATT(c_i)=\{a_1, a_2,..., a_n\}$ (*e.g.*, name, definition, synonym, *etc*.). We define a set of concepts of an ontology O_X at time *j* as $CNC(O_X^j) = \{c_1^j, c_2^j,..., c_n^j\}$. A relationship *r* interconnects two concepts and has a specific type, *e.g.*, "subsumption", "part-of", *etc*.

The context CT of a concept c_i in the ontology stands for the union of the sets of $sup(c_i)$, $sub(c_i)$ and $sib(c_i)$ concepts of c_i, as following:

$$CT(c_i) = sup(c_i) \cup sub(c_i) \cup sib(c_i)$$

where :$sup(c_i) = \{c_k | c_k \in CNC(O), c_i \sqsubset c_k \wedge c_i \neq c_k\}$, $sub(c_i) = \{c_k | c_k \in CNC(O), c_k \sqsubset c_i \wedge c_i \neq c_k\}$, $sib(c_i) = \{c_k | c_k \in CNC(O), sup(c_k) \cap sup(c_i) \neq \varnothing\}$ to which $c_i \sqsubset c_k$ stands for "*ci* is narrower or more specific than *cj*", *e.g.*, "*hypotension*" is more specific than "*vascular disease*".

An ontology mapping $M_{S,T}^j$, established at time *j*, interrelates a set of given concepts by semantic correspondences. A correspondence $cor_{st} = (c_s, c_t, conf, semType)$ links two concepts $c_s \in CNC(O_S^j)$ and $c_t \in CNC(O_T^j)$ through the *semType* relation. The *conf* denotes the similarity value between c_s and c_t indicating the confidence of their relation. We consider the following types of semantic relation: *unmappable* [⊥], *equivalent* [=], *narrow-to-broad* [≤], *broad-to-narrow* [≥] and *overlapped* [≈].

We have expressed behaviours of mapping adaptation as ***Mapping Adaptation Actions*** [2] (*cf.* Figure 1). In *MoveM* the source concept c_s of the correspondence is replaced by another concept c_k. Similarly, in *DeriveM* the original correspondence remains and a new correspondence appears connecting a concept c_k with c_t. In both actions, $c_k^1 \in CT(c_s^1)$. In *RemoveM* the correspondence is deleted and the action *ModifySemTypeM* consists in modifying the type of semantic relation. The *no-action* refers to the cases where correspondences remain unchanged.

We conduct our experiments using various releases of official mappings established between biomedical ontologies including: SNOMED-CT (SCT), MeSH, ICD-9-CM (ICD9) and ICD-10-CM (ICD10). In particular, mappings interconnect SCT(2010)-ICD9(2009) and SCT(2012)-ICD9(2011); MeSH(2012)-ICD10(2011) and MeSH(2013)-ICD10(2011). We observe the evolution of the biomedical ontologies and mappings through the following procedure:

- **Calculate MAAs:** Given all correspondences between two ontologies, for each correspondence impacted by some ontology change, we determine a list of MAAs. We remove correspondences where the source concept remains unchanged from one ontology version to another, or where ontology changes simultaneously affect both source and target concepts. We use the *COnto-Diff* tool to calculate ontology changes [3]. Having successive releases of mappings allows us to calculate the MAAs for each correspondence. We compare the elements composing a correspondence (identifier of source and target concept, and relation type). For instance, given a correspondence at time *j*, we search its elements in the mapping

at time *j+1*. If we fail to find, we sign the *RemoveM* action for such correspondence. We use a similar approach to determine the other MAAs.

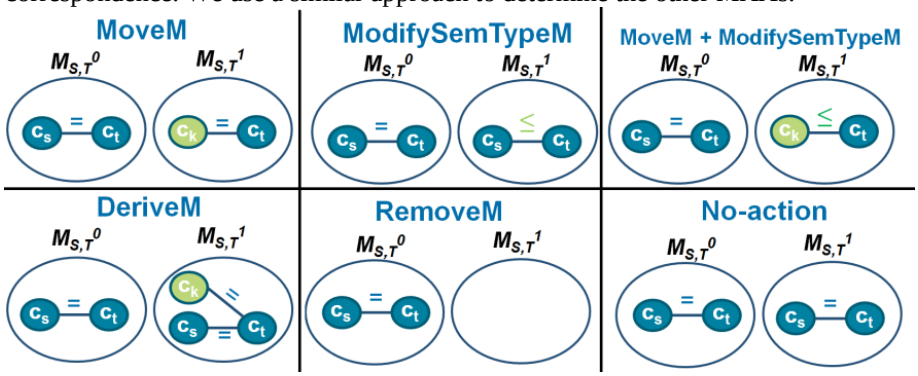

Figure 1. Mapping Adaptation Actions

- **Identify relevant attributes:** For each source concept impacted by ontology changes, we identify a minimal set of relevant attributes (we assign 3, but it can be parameterized) representing the most relevant attributes for a given correspondence, since this is established on partial information (*e.g.*, some attributes) defining concepts. These source concept attributes consist of the most similar to the ones in the target concept. We pre-selected groups of attributes to calculate the similarity value. For instance, we do not compare the ICD9 attribute "Exclude" with the SCT attribute "Name", if a correspondence exists.
- **Calculate similarity:** For each relevant attribute of a source concept c_s in an impacted correspondences at time j_0, we calculate the similarity relatedness with attributes issued from concepts in the context of c_s at time j_1. To this end, we explore string-based similarity metrics, especially the *bi-gram* because this measure performs well on ontology matching tasks [7]. The similarity function receives two attribute values and returns a value ranging from 0 to 1. The higher the result is, the more similar these attributes are. We analyze the density of calculated similarity values ranging from 0 to 1 for each MAAs by using the kernel density estimation method [8]. The density stands for a smoothing distribution of frequencies of the similarity values similar to a histogram. We observe the similarity's influence on the MAAs by studying the following scenarios:
  1. **REL_ATTS_changedCT:** For each impacted correspondence, we search the highest similarity value calculated among the relevant attributes identified with all attributes issued of changed concepts in the context of c_s at time j_1. We register such maximum similarity value for the MAAs applied.
  2. **Best_REL_ATT_changedCT:** This is similar to scenario 1, but we only consider the best relevant attribute for calculating the similarity (*i.e.*, the most significant attribute for c_s regarding the c_t). We aim to examine the influence of this particular attribute with respect to all relevant attributes.
  3. **REL_ATTS_unchangedCT:** This scenario also performs similar to scenario 1, but examines the unchanged part of the context of c_s (*i.e.*, all unchanged concepts in the context).
  4. **Best_REL_ATT_unchangedCT:** Similar to scenario 2, we observe the similarity density only considering the best relevant attribute with the unchanged context.
  5. **REL_ATTS_conceptMAA:** This scenario applies only for *MoveM* and *DeriveM* since they involve a concept c_k^1 (denoted *conceptMAA*) which

replaces the concept c_s. We calculate the density of the highest similarity values among the relevant attributes and the attributes characterizing c_k^1.

6. **Best_REL_ATT_conceptMAA:** This is similar to scenario 5, but we only consider the best relevant attribute.

We repeated this procedure for both set of mappings (SCT-ICD9 and MeSH-ICD10). The total number for each type of MAAs analyzed remained: *MoveM* (635); *DeriveM* (768); *ModSemTypeM* (58), *RemoveM* (193) and *no-action* (9280).

3. Results

Figure 2 presents the obtained kernel estimation based density of the similarity values for the *DeriveM* and *MoveM* actions. For *DeriveM*, we observe a similar behaviour of the density of similarity values in scenarios 1 (**REL_ATTS_changedCT**) and 2 (**Best_ REL_ATT__changedCT**), revealing a high frequency of low similarity values. We explain this by the fact that most frequently the changed CT set is empty so similarity values remain zero. In scenarios 3 (**REL_ATTS_unchangedCT**) and 4 (**Best_ REL_ATT_unchangedCT**), we observe the contrary behaviour with a bigger density of high similarity values. This shows that the original source concepts appear similar with unchanged concepts in the context. Scenario 4 presents a lower density of high similarity values compared with scenario 3. We point out a higher density of medium similarity in scenario 5 (**REL_ATTS_conceptMAA**) compared with scenario 6. This reveals that the best relevant attributes frequently are not the most similar ones to *conceptMAA*. Results for scenarios in *MoveM* indicate a similar behaviour of the *DeriveM* except for the scenario 1 (**REL_ATTS_changedCT**). This underscores that when finding similarity with the changed CT a *MoveM* action occurs.

Figure 3 presents the achieved results for the *ModSemTypeM*, *RemoveM* and *no-action*. For the *ModSemTypeM*, scenarios 1 (**REL_ATTS_changedCT**) and 2 (**Best REL_ATT_changedCT**) behave very similar, with a high density of low similarity values. On the contrary, scenario 3 (**REL_ATTS_unchangedCT**) shows a higher density of high similarity values which remains lower for the scenario 4. All scenarios for *RemoveM* behave very similar presenting a high density of low similarity values. This may rely on the fact that when correspondences evolve by *RemoveM* action, the whole context contained unchanged or not similar concepts that could be candidates for replacing the source concept. When *no-action* is applied, the scenarios with the changed context (1 and 2) show a high density of low similarity values also because of the inexistence of changed concepts. However, we observe a high density of high similarity values in the scenarios with unchanged context (3 and 4) despite that correspondences remained unchanged (*i.e.* any mapping adaptation action was applied).

4. Discussion

Our results pointed out well-defined behaviours with respect to the density of similarity values for the MAAs. We found the similarity between ontology versions a relevant factor for mapping adaptation. This might help deciding which action applying, but seems insufficient for a completely automatic mapping adaptation. For example, for *DeriveM* and *MoveM* actions (*cf.* Figure 2), the density of the highest similarity values in scenarios 3 and 4 (unchanged context) is bigger than the density observed in scenarios 5 and 6 (*conceptMAA*). We expected that the similarity value with the concept where a *DeriveM* or *MoveM* action happens would remain higher than the similarity found with unchanged concepts. On the other hand, the similarity aspect can help deciding for a *RemoveM* action (*cf.* Figure 3). In addition, we expected to observe

a more determinant influence of the best relevant attributes in the studied scenarios. We conclude that a system for mapping adaptation must combine the similarity with other aspects to achieve enough facts enabling to trigger different MAAs. Our future work involves studying these further factors.

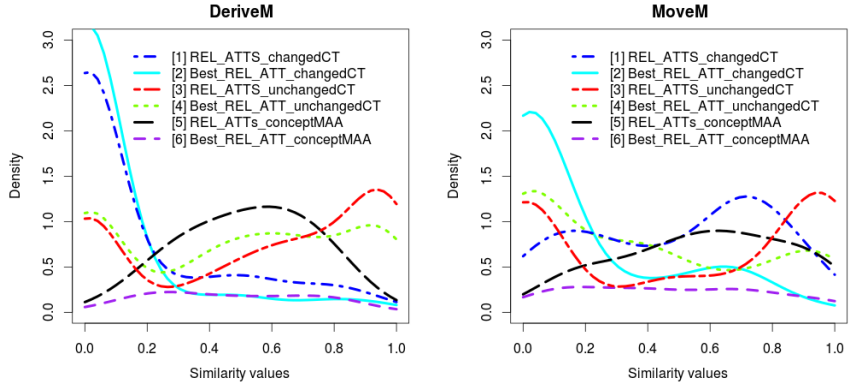

Figure 2. Similarity values density between concept attributes for *DeriveM* and *MoveM* actions

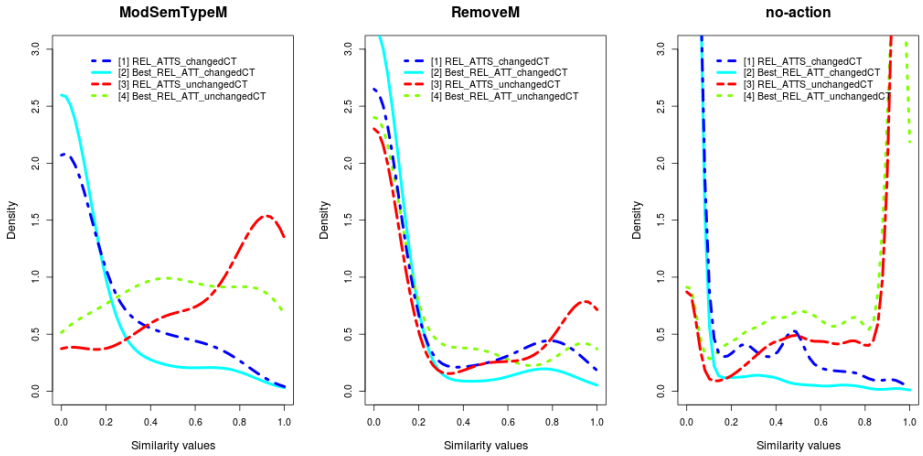

Figure 3. Similarity value density between concept attributes for *ModifySemTypeM, RemoveM* and *no-action*

References

1. Burgun, A., Bodenreider, O.: Accessing and integrating data and knowledge for biomedical research. *Yearbook of medical informatics*. (2008), 91–101.
2. Dos Reis, J.C., Dinh, D., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Mapping adaptation actions for the automatic reconciliation of dynamic ontologies. *CIKM*. (2013), pp. 599–608.
3. Hartung, M., Groß, A., Rahm, E.: COnto-Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. *Journal of biomedical informatics.* 46 (2013), 15–32.
4. Gross, A., Hartung, M., Thor, A., Rahm, E.: How do computed ontology mappings evolve ? A case study for life science ontologies. *IWOD workshop at ISWC*. (2012), pp. 1–12.
5. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Understanding semantic mapping evolution by observing changes in biomedical ontologies. *JBI*. 47 (2014), 71-82.
6. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Characterizing Semantic Mappings Adaptation via Biomedical KOS Evolution: A Case Study Investigating SNOMED CT and ICD. *AMIA Symposium.* (2013), pp. 333–342.
7. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. *ISWC*. (2013), pp. 294–309.
8. Sheather, S.J., Jones, M.C.: A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Royal Statistical Society*. 53 (1991), 683–690.